Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 8
Open Access

A Longitudinal And Domain Wise Survey On MIMIC III Dataset

[1]Ch. Raja Ramesh, [*2] Pantula Muralidhar, [3]K. M. V. Madan Kumar, [4]Srinu Banothu, [5]Manoj Kumar Mahto, [6] R. Chaitanya Kumar

[1]Associate Professor, Computer Science & Engineering (DS), Vignan Institute of Technology & Science, Hyderabad. Email: chrajaramesh@gmail.com

[*2]Associate Professor, Computer Science & Engineering, Vignan Institute of Technology & Science, Hyderabad. Email: vijayamuralisarma@gmail.com- Corresponding Author

[3]Professor, Computer Science & Engineering, Vignan Institute of Technology & Science, Hyderabad. Email: madankukunuri@gmail.com

[4]Associate Professor, Computer Science & Engineering (DS), Vignan Institute of Technology & Science, Hyderabad. Email: srinub1307@gmail.com

[5]Associate Professor, Computer Science & Engineering (DS), Vignan Institute of Technology & Science, Hyderabad. Email: manojkr.bit@gmail.com

[6]Assistant Professor, Computer Science & Engineering, Vignan Institute of Technology & Science, Hyderabad. Email: rchkumar86@gmail.com

Cite this paper as: Ch. Raja Ramesh, Pantula Muralidhar, K. M. V. Madan Kumar, Srinu Banothu, Manoj Kumar Mahto, R. Chaitanya Kumar, (2024) A Longitudinal And Domain Wise Survey On MIMIC III Dataset. *Frontiers in Health Informatics*, *Vol.13 No.08*,

ABSTRACT

Medical Information Mart for Intensive Care III (MIMIC-III) database is one of the popular databases in medical fields related to the history of in-and-out patients. This database is applied in multiple domains such as data analysis, Machine learning, and Natural Language processing to study or analyze the characteristics of the patients deceased or suffered with sepsis. To explore the extensions with MIMIC database, the most popular articles are scrutinized then identified overview of the concepts related to various computer science domains, summarizes the architecture models, optimization models, and statistical measures of different features used from the MIMIC database, and compared the among articles related to the same class of diseases. The findings, provided in this article open doors for future scope in research on MIMIC Database that will be helpful for different medical fields as well as in the fields of computer-related domains.

Keywords: MIMIC Database, Sepsis, Machine Learning, Natural Language Processing etc.

INTRODUCTION

The recent advancements in medical field research are possible because of the availability of databases having patient history, one such database is MIMIC-III. The MIMIC-III database includes a wide range of data features related to patients admitted to critical care units. Using these data features various scientific studies are conducted such as predicting 30-day mortality of the patients, predicting sepsis, mortality prediction for cancer patients as well as Trauma patients, and lung cancer stay prediction. These studies support the necessity of the growth in medical fields to improve patient treatments. These data features are not only limited to analyzing the patient conditions but also aid in developing models for predicting the patients using computer science domains such as Machine Learning, Deep Learning, and Natural language processing.

The major data features used in the MIMIC-III database for popular case studies are discussed in figure 2. These features are the primitives for applying scientific analysis on MIMIC database. These are the primitives that provide a comprehensive view of the clinical care provided to patients in the critical care units.

- Demographics: Information such as inpatient and dates and discharge dates, birth date or date of death, religion, ethnicity, and marital status.

- Monitoring Every second: waveforms, Vital signs, trends, and alarms.
- Chart Data: Details on fluids, medications, and progress notes.
- Notes and Reports: Discharge summaries, radiology reports (X-ray, CT, MRI, Ultrasound), and cardiology reports (ECHO, ECG).
- Orders: Provider order entry (POE) data.
- Laboratory Tests: Results of laboratory tests.
- Microbiology: Microbiology test results.
- Billing Information: Including ICD9 codes, DRG codes, and procedure codes (CPT).
- Social Security Death Index Data.
- User Feedback and Corrections.

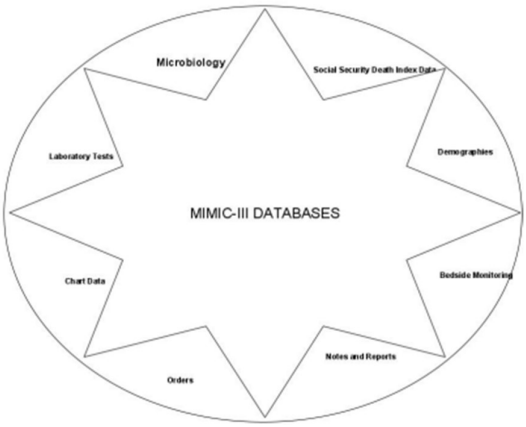


Figure 1 Primitives and data features of MIMIC database

The key features of the database including vital signs, medications, hospital length of stay, observations, notes, procedure codes, diagnostic codes, imaging reports, laboratory measurements and survival data provide a holistic view of patient care in critical care units and aid in support for the research on various aspects of critical care, such as patient outcomes, treatment effectiveness, disease progression, and healthcare interventions. The data enables researchers to analyze trends, patterns, and correlations to generate new insights and knowledge.

By analyzing data on patient outcomes, treatment protocols, and clinical practices, healthcare teams can identify areas for improvement and implement evidence-based strategies to enhance patient care and safety.

MIMIC-III also serves as a valuable educational resource for teaching and learning in healthcare analytics, data science, and clinical research. Educators can use the database to provide students with real-world clinical data for analysis, research projects, and coursework, enhancing their understanding of critical care practices and data analysis techniques. MIMIC-III promotes producible and transparency in research by providing a standardized dataset that can be accessed and analyzed by multiple researchers. This transparency enhances the credibility and reliability of research findings and encourages scientific collaboration and validation.

Overall, the MIMIC-III database offers a robust platform for critical care analysis, research, quality improvement, education, and international collaboration, contributing to advancements in critical care practices and patient outcomes. This also creates benefits for the researchers because rich source of de-identified clinical data which aids in exploring topics such as machine learning approaches for predicting patient outcomes, clinical decision support systems, and epidemiological studies. By studying patterns in patient data, healthcare professionals can identify areas for improvement and implement strategies to enhance patient outcomes.

Therefore, the MIMIC-III database serves as a valuable resource for advancing research, improving healthcare practices, and enhancing education in the field of critical care and clinical data analysis. Because of the broad scope of the MIMIC-III database, this research article explores recent articles and forecasts the large-scale scope of critical analysis in MIMIC-III, by considering the following

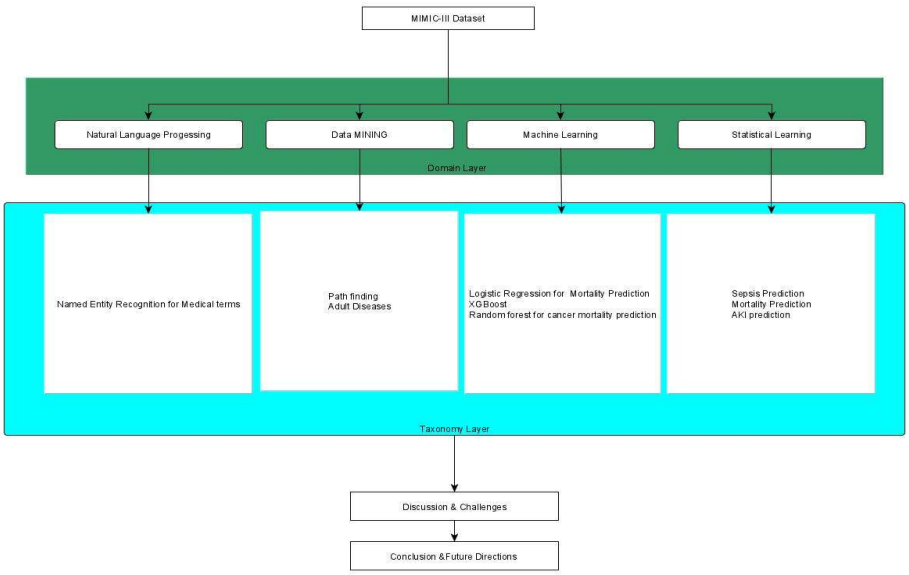
2. OBJECTIVES

1. Gathering the information from the recent articles about various domain areas using the MIMIC-III database and exploring the similarities of the articles under the identical domains.

2. Identifying and analyzing the feature requirements and models for improvement of statistical investigations and medical diagnosis.

3. ARCHITECTURE

In achieving the major objectives of this article, the author considered the architecture shown in figure 1. The research papers from 2016 to 2024 are considered using the MIMIC-III database have been partitioned into various domains such as Data Analysis, Natural language processing, machine learning, and deep learning and it was further categorized into sub-topics like Named Entity Recognition, Mortality prediction, cancer patient stay prediction, etc.



Even though, MIMIC-III provides a vast amount of data related to patients' critical care. Since the dataset was partially public i.e., must need permission to access the database, which was a major drawback for lagging in the research, which created a key challenge to summarizing the research papers related to the MIMIC-III database.

This article has made summarization on many aspects of the MIMIC-III such as predicting mortality, predicting trauma patients' stay, etc. by focusing on brief literature to provide the future research scope on MIMIC-III database.

4. STATISTICAL ANALYSIS ON MIMIC-III DATABASE:

Figure 2: Article Flow Architecture

4.1 Sepsis:

The author Dan-Ni et.al., [2] examines the association between Body Mass index (BMI) and the patients discharged after affecting with sepsis. This article considers different BMI categories under weight, normal, overweight and Obese. The experimentation in the article includes logistic regression or Cox Proportional hazard model. The key findings of the article includes overweight and obese patient tends slow better and survivals than underweight and normal weight.

4.2 Cardiac surgery:

Yiran zhang et. al., [22] claims that when comparing overweight and normal weight persons after the cardiac surgery in the elderly persons, the overweights patients had better 1-year survival rates. He considers patients with open heart surgery with ECC (ExtraCorporeal Circulation) (ICD-9 code: 3961), and considered population into four groups based on BMI values. They made the statistical analysis called (MLRA) Multivariate Logistic Regression Analysis on the parameters such as ventilation duration and length of hospital stay when compared the 1 year mortality rate is shorter in over weights patients compared with normal weight persons.

Dongliang Liu et. al., [15] showed interest on Subarachnoid hemorrhage (SAH), a high mortality type stroke disease, by predicting hyperglycemia among the critically ill patients outcomes very poor clinical results.

The author analyzed 30-day and 90-day mortality among the critically ill patients with SAH. Generalized additive model used to identify the relation between the blood glucose and 30 and 90 day mortality of the ill patients.

4.3 Acute kidney Injury(AKI):

Caijje Liu et. al., [14] analyzed the relation between short-term prognosis and Transthoracic echocardiography (TTE) in ICU patients. According to the author, there are 10-15% of population are having problem of AKI. For the statistical analysis, population is categorized into TTE (Transthoracic echocardiography) and non TTE groups where 28-day mortality was more for non TTE groups and claimed that TTE examination may protect the mortality for patients with AKI in the ICU. GBM (Gradient Boost Model) used to measure the PS (propensity score) of the patients for TTE examination.

Didi Han et. al., [7] discusses about Severe acute pancreatitis (SAP) which would be the major cause of death for most of the ICU admitted patients. The author analyzed the data using multivariate Cox regression on the nomogram evaluated by the Harrell's concordance index. The author presented the 28-day, 60-day and 90-day mortality rates based on the nomograms.

4.4 Cancer:

Among the many cancers bone marrow is having low survival and high morbidity rates. Guiqiang Miao et.al., claims to predict the morbidity risk of the patients by validating the nomograms [17]. For this the dataset was considered 70\% as training cohort and 30\% as validation cohort and was analyzed using regression methods such as Univariate and multivariate Cox regression analysis. From the nomogram 30-day morbidity rate among patients was predicted.

5. DATA MINING AND ANALYSIS STRATEGIES ON MIMIC-III DATASET

5.1 Cancer:

Angelina Prima Kurniati et. al., [12] uses process mining create pathway for the different types of cancer diseases. The author uses three steps to obtain the path where step 1 deals with preprocessing the data, step 2 discuss about establishing control flow models and step 3 intends for integrating the process models.

5.2 Adult disease characteristics and mortality

A research article identified the steps involved in The data analysis and extraction to determine disease distribution among adult patients in the MIMIC-III database. This articles selected only the adult patient's analysis, and unique patient levels were considered. The last admission record for each patient was chosen, and the last ICU record in the last admission record was selected for statistical analysis. The first diagnosis was selected as the object of data analysis, and disease classification was refined based on primary International Classification of Diseases (ICD-9) codes. Hospitalization days were calculated by subtracting the admission time from the discharge time for each patient. Analysis of Mortality calculated based on date of death for each patient was recorded in the database. The number of days a patient died after discharge was calculated by subtracting the date of death from the date of discharge. Patients who died within 90 days were selected for further analysis. The prevalence of the top 10 diseases with the largest number of patients was calculated, and the distribution of these diseases in different ICUs was analyzed to understand the main diseases in each ICU more clearly. The highest mortality rate observed among patients in the study was associated with Septicemia, with a mortality rate of 48.9\%. Septicemia had the highest mortality rate among the top 10 diseases with the largest number of patients in the study. Following Septicemia, Intracerebral hemorrhage and Other diseases of lung also had high mortality rates exceeding 44\%. Patients with Other diseases of lung had the longest ICU stay among the top 10 diseases analyzed.

6. MACHINE LEARNING MODELS ON MIMIC-III DATASET

6.1 Sepsis:

As per statistics of public health globally about 30% of the population dies because of sepsis annually in the ICU [20]. Sepsis is the abnormality disease induced by the infection which becomes a life treat to the patients [16]. Due to complex Syndrome and it is necessary for prior detection of Sepsis in ICU patients has become one of the major objective for several researchers. Early detection of Sepsis can help to provide appropriate treatment for the patients to improve the survival rate of the Sepsis patients. Thankful to the growth of machine learning technologies for predicting these type of diseases in the early scenarios. MIMIC-III database has become popular to train the models with state-of-the-art benchmark accuracies. The Sepsis prediction using machine learning is majorily focused on temporal division of patients admission in ICU.

An article [8] deals with predicting the mortality rate of ICU patients with in the one-month of admission with life threat disease. In this process they identified patient affected with the sepsis. They have taken the ICU data from the MIMIC III database and divided in two types: survival and deceased patients in last 30 days of admission. XGboost machine learning approach, following variables such as respiratory rate, temperature, oxygen saturation, anion gap, creatinine levels, hemoglobin levels, lactate levels, potassium levels, sodium levels, blood urea nitrogen levels, white blood cell counts, international normalized ratio, urine output, score system, comorbidity, and sources of infection were selected based on clinical significance and availability. These variables were then compared between groups of patients who survived or died within 30 days.

An article [19] identifies the number of patients affected with the sepsis decease after admitting with the span of 3 to 5 hours and at the time of admission patients are not having the sepsis disease. This article proposes (RNN) recurrent neural network model for the prediction of sepsis onset. This will provide the time dependent patterns with in the dataset followed by sepsis onset. In RNN they have taken gated recurrent unit (GRU) as a hidden layer architecture. Binary cross-entropy cost to cost function optimizes the using this GRU network network. By conclusion of this article predicting the gradual development of sepsis disease after admitting in ICU by using RNN. They are also show that further research process is necessary to identify the proper sepsis onset detection as it varies depending on the amount of accepted interpolations.

In this paper [4] the authors aim for early detection of sepsis by proposing a pipeline approach that includes data extraction, data enhancement, augmented feature generation, and transfer learning. Medical research often faces issues of data sparsity and imbalance. The authors address these by using self-attention mechanisms to identify correlations within unlabeled data. They introduce an interpretable feature augmentation method that captures potential correlations between features, creating generic representations for missing data parts by leveraging high correlations between different histological data. This provides a more complete view of the patient's condition. Furthermore, they employ conditional transfer learning and a data blending approach, which transfers knowledge from source data using a small amount of target data, overcoming the limitations of adapting augmentation methods to varying data distributions from different sources.

6.2 Cancer disease:

In a paper 'CanICU' [9] demonstrated high sensitivity and specificity in predicting 28-day mortality in critically ill cancer patients, outperforming the SOFA score in the validation cohort. The CanICU model showed robust performance in external validation cohorts, indicating its potential for generalizability and reliability. Highlights of this work is the potential of 'CanICU' as a valuable tool for predicting short-term mortality in critically ill cancer patient. The model harnesses electronic patient records from the first 24 hours before ICU admission to predict short-term and long-term mortality. The CanICU model demonstrated high predictive accuracy in determining the 28-day mortality of critically ill cancer patients. By utilizing nine easily obtainable variables, the model outperformed traditional scoring systems and showed promising results in mortality prediction. This study also conducted external validation of the CanICU model using data from multiple medical institutions, including the Yonsei Cancer Center and Samsung Medical Center in Korea, as well as the Medical Information Mart for Intensive Care (MIMIC) database in the USA. The model's performance was validated across different patient cohorts, enhancing its generalizability and reliability.

B Alsinglawi et. al., [1] designed a machine learning framework for predicting lung cancer study gives the hospital length of stay. This work introduces a predictive framework using supervised ML methods with the Random Forest model achieving the best results. By utilizing over-sampling and under-sampling techniques, we were able to achieve high AUC results and explain the model outcomes using the SHAP technique. The Random Forest model's superior performance, robustness, and consistency in handling feature selection methods and predicting outcomes contributed to its outperformance compared to other models in predicting lung cancer inpatients LOS. By using over-sampling and under-sampling methods in dealing with imbalanced datasets can lead to improved model performance, enhanced predictive power, reduced bias, increased sensitivity and specificity, and optimized AUC scores, ultimately resulting in more accurate and reliable predictions in classification-based approaches. The SHAP technique aids in providing local explanations, enhancing interpretability, offering clinical insights, promoting explainable AI, and guiding clinical information systems by explaining the outcomes of the predictive model in the study on lung cancer LOS prediction. This study provides valuable insights into predicting lung cancer inpatients LOS using machine learning models, emphasizes the importance of addressing imbalanced datasets, and highlights the significance of explainable AI techniques like SHAP for model interpretability. Future enhancements could focus on external validation, analyzing larger datasets, hyperparameter tuning, and ensuring robust predictive outcomes in real-world healthcare settings.

The physiognomies and clinical subtypes of cancer patients in the severe care unit are studied in a research paper [5]. Altered clinical subtypes of ICU patients with cancer were well identified based on type of admission and clinical service provider using K-means clustering. The study highlighted the importance of caution when considering cancer patients in the ICU as a whole population, emphasizing the need to recognize the diversity among these patients in both clinical practice and research. The APACHE IV counting system was found to perform better than the SAPS II system for cancer patients at low-slung risk of mortality in the ICU. The study suggests that further validation is needed to determine the performance of these scoring systems for cancer patients at high risk of mortality in the ICU. These outcomes provide valuable insights into the complexity of managing cancer patients in the ICU and underscore the significance of personalized approaches based on clinical subtypes and risk assessment tools. To get all the above outcomes used several algorithms, In this K-means clustering algorithm was used to recognize unlike clinical subtypes of ICU patients with cancer based on admittance type and clinical service provider. Logistic regression algorithm was used for imputing absent values of variables fed into the K-means model. Principal component analysis (PCA) algorithm was used to compress all variables of

each patient into two components for visualization on a scatterplot. Additionally, this study refers to the APACHE IV system, which is not an algorithm but a severity scoring system used for assessing the severity of illness in ICU patients. These algorithms and systems were instrumental in analyzing the data and identifying distinct clinical subtypes of cancer patients in the intensive care unit. The outcome of the paper provides valuable insights into the outcomes of lung cancer patients admitted to the ICU. The study found that the median survival of these patients was 2.93 months, with a 28-day in-hospital death rate of 30.6\% and a 6-month death rate of 68.2\%. Factors such as age, illness ruthlessness scores (SAPS II and SOFA), metastatic status, and the use of mechanical ventilation were identified as independent risk factors for mortality.

In another paper [21], A model demonstrated that the XGBoost algorithm outperformed nine other machine learning algorithms in predicting 90-day mortality for ICU trauma patients. The model identified ten key features strongly associated with 90-day mortality in these patients. These features are crucial for predicting patient outcomes and mortality risk. The study underscored the effectiveness of the XGBoost algorithm in predicting mortality outcomes for ICU trauma patients, providing valuable insights for clinicians to identify high-risk patients and implement early interventions. The model did acknowledge limitations, including the retrospective nature of the analysis, potential bias in the patient population, and the need to explore additional features in the dataset. Future research should focus on addressing these limitations and further validating the predictive model. The primary goal of this retrospective cohort study was to develop and validate a predictive model using the XGBoost algorithm that outperformed other machine learning algorithms in predicting 90-day mortality outcomes.

In another study in based on other paper [3], the focus was on patients with thrombosis or cancer, aiming to develop and validate interpretable machine learning models for predicting early and late mortality. The predictive models for early mortality showed excellent performance in both disease categories, with high area under the receiver operating characteristic curve (AUC-ROC) values: VTE-MIMIC-III 0.93, eICU 0.87, cancer-MIMIC-III 0.94. These models were externally validated and compared with traditional scoring systems, demonstrating superior performance in predicting both early and late mortality outcomes. The main machine learning algorithms used in the study were Random Forest (RF) and XGBoost. RF was used for training the models, with parameters such as training 500 trees using the deviance splitting criterion and a minimum leaf size of 3. The XGBoost algorithm was employed to develop predictive models and address class imbalance in the datasets. These algorithms were selected for their effectiveness in handling complex datasets, addressing class imbalance issues, and providing accurate predictions for early and late mortality outcomes in critically ill patients with venous thromboembolism.

6.3 Kidney Disease:

Gao Wenpeng et. al., [6] proposed a decision tree model called lightgbm for predicting acute kidney injury in ICU patients. The study uses physiological and biochemical indicators of patients and construct the models of logistic regression, random forest and lightgbm for 24h, 48h and 72 h respectively and predict the kidney disease for ICU patients.

6.4 Heart failure:

Fuhai Li et. al. [13] predicts the in-hospital mortality for ICU patients uses gradient boosting (XGBoost) and least absolute shrinkage and selection operator regression models LOSSO regression models. In comparison author identified XGBoost model as the final model to build nomogram and the nomograms are best predictors for the predicting the mortality.

7. NATURAL LANGUAGE PROCESSING USING MIMIC-III DATABASE

7.1 Annotations tools:

The author [10] proposes biomedical NER based on BiLSTM-CNN-Char DL architecture. The Named Entity Recognition, which is a technique of NLP establish new state of art results with best accuracy compared to other well known biomedical NER's.

Edward Moseley et. al., [18] analysis the patient Electronic Health Records(EHR) to retrieve the study the patients' conditions and treatment. From the EHR developed a dataset termed patient phenotyping.

Kraljevic et al., [11] develops a entity recognition and linking model called Medical Concept Annotation tool(MEDCAT), it can extract structure and information from the EHR. The author claims that MEDCAT is best for disease detection from EHR. MEDCAT is deployed in many hospitals in real time.

8. DISCUSSION AND CHALLENGES

The MIMIC-III database is a popular educational resource in the medical field. It deals with discrete fields of the patients who admitted in the hospitals such as Demographics, vitalsigns, medications, hospital length of stay, observations, notes, procedure codes, diagnostic codes, imaging reports, laboratory measurements and survival data. It provides a lot of features that can automate and simply the detection of prediction of the complex medical diseases of the patients. EHR of the patients are also aid to study the hidden pattern to identify the complex

unpredictable diseases.

As a study, there are lot amount of techniques are referred to identify the challenges in the medical field of which MIMIC-III has provided vast information about the medical records of the patients. One major disadvantage was it was half public and to maintain privacy the records data was modified intentionally. Due to which the exact information relevant to the patient was hidden and at the same time some of the pattern which are based on the demographic details are truncated.

This study explores a deep survey on MIMIC database-III and categorized the research into various domains such as statistics, machine learning, data mining and natural language processing, Of which statistics and machine learning techniques are very much close to each other. Nomograms are identified as important features in most of the diseases in statistical analysis and machine learning. Datamining is majorly used for extraction of complex patterns in the provided details, while natural language processing extract the valuable information from the medical discharge summaries.

Figure 1 explores the follow this article majorly focused on the topics of sepsis, AKI, cancer detection of the patients to predict the mortality rate in 28, 60, and 90 days. Eventhough a lot of techniques were developed on the database the major pitfall was amount of data available was lumbersome related to all category of diseases but when consider for a single disease it was limited. The major challenges observed in the database was curse of dimensionality, Accessibility, Timeliness, completeness, format of usability, Transperancy and Access of Control.

Curse of Dimensionality: The databases consists of vast amount of data related to each patient such as Demographics,vital signs, medications,hospital length of stay, observations, notes, procedure codes, diagnostic codes, imaging reports,laboratory measurements and survival data. This creates ambiguity in predicting for particular use case.

- Accessibility: It is hard for the researchers until they have proper medical terminology in all the aspects of hospital information.
- Timeliness: Inorder to provide privacy the data is garbled in different manner makes difficult to identity essential entries of the patients, and loss some minute valuable information.
- Completeness: The database is not provided with continuous updations some makes difficult for identification of analysing regular followups for the patients.
- Format of Usability: The data is available with multiple format examples admission details are in excel files and discharge summary in seperate text files.
- Transperancy: Missing clear information about the patients inorder to provide privacy of the patients.
- Access of Control: The database was semi public, inorder to access the database the person have to pass certain criteria's. For a large domain of researchers in the medical field also the MIMIC database is unknown.
- Exploration: The models developed in many cases are very old such as LOSSO logistic regression, decision trees etc. Wide improvement must be focused on the database with the techniques using Convolutional networks, recurrent networks, GAN networks, attention mechanisms, and capsule networks etc.

9. CONCLUSION & FUTURE DIRECTIONS

The survey has studied the articles that are publicily available on the google scholar in the period of 2016 to 2024 related to different domains such as statistical analysis, machine learning, natural language processing and datamining on predict the diseases like sepsis, SKI and cancer. MIMIC-III having the data for the period of 2001 to 2012 with various details including admissions, patients, ICU stays, and various clinical data tables of more than 40,000 unique patients with 60000 data which is vary small when multiple domains of medical fields are considered. As a study this article Gathers the information from the recent articles using the MIMIC-III database about various domain areas and observed the similarities in techniques used in several disease predictions and explored the pitfalls in MIMIC-III database.

As a future direction the relevant work can also be compared with MIMIC -IV database, which is having patients information from 2008 to 2019, which having normalized schemas and 380000 volumes of data available

REFERENCES

1. Belal Alsinglawi, Osama Alshari, Mohammed Alorjani, Omar Mubin, Fady Alnajjar, Mauricio Novoa, and Omar Darwish. An explainable machine learning framework for lung cancer hospital length of stay prediction. Scientific reports,12(1):607, 2022.
2. WNAG Dan-Ni, HUA Li-Dian, PAN Zhi-Guo, WEN Qiang, and SU Lei. The impact of body mass index on the prognosis in sepsis patients: A retrospective analysis on account of the large clinical database

mimic-. Jie Fang Jun Yi Xue Za Zhi, 46(2):129, 2021.

3. Vasiliki Danilatou, Stylianos Nikolakakis, Despoina Antonakaki, Christos Tzagkarakis,Dimitrios Mavroidis, Theodoros Kostoulas, and Sotirios Ioannidis. Outcome prediction in critically-ill patients with venous thromboembolism and/or cancer using machine learning algorithms: external validation and comparison with scoring systems. International Journal of Molecular Sciences, 23(13):7132, 2022.
4. Yutao Dou. Biomarkers for Early Detection of Sepsis: A Multi-View Machine Learning Approach. PhD thesis, 2022.
5. Shaowei Gao, YaqingWang, Lu Yang, ZhongxingWang, andWenqi Huang. Characteristics and clinical subtypes of cancer patients in the intensive care unit: a retrospective observational study for two large databases. Annals of Translational Medicine, 9(1), 2021.
6. Wenpeng Gao, Haijin Lyu, Lang Zhou, and Shengwen Guo. Decision tree algorithm applied to mimic-iii database for the prediction of acute kidney injury in icu patients. Urinary and Renal Research, 2(1):2025, 2021.
7. Didi Han, Fengshuo Xu, Chengzhuo Li, Luming Zhang, Rui Yang, Shuai Zheng, ZichenWang, Jun Lyu, et al. A novel nomogram for predicting survival in patients with severe acute pancreatitis: an analysis based on the large mimic-iii clinical database. Emergency Medicine International, 2021, 2021.
8. Nianzong Hou, Mingzhe Li, Lu He, Bing Xie, Lin Wang, Rumin Zhang, Yong Yu, Xiaodong Sun, Zhengsheng Pan, and Kai Wang. Predicting 30-days mortality for mimic-iii patients with sepsis-3: a machine learning approach using xgboost. Journal of translational medicine, 18:1–14, 2020.
9. Ryoung-Eun Ko, Jaehyeong Cho, Min-Kyue Shin, Sung Woo Oh, Yeonchan Seong, Jeongseok Jeon, Kyeongman Jeon, Soonmyung Paik, Joon Seok Lim, Sang Joon Shin, et al. Machine learning-based mortality prediction model for critically ill cancer patients admitted to the intensive care unit (canicu). Cancers, 15(3):569, 2023.
10. Veysel Kocaman and David Talby. Accurate clinical and biomedical named entity recognition at scale. Software Impacts, 13:100373, 2022.
11. Zeljko Kraljevic, Daniel Bean, Aurelie Mascio, Lukasz Roguski, Amos Folarin, Angus Roberts, Rebecca Bendayan, and Richard Dobson. Medcat–medical concept annotation tool. arXiv preprint arXiv:1912.10166, 2019.
12. Angelina Prima Kurniati, Geoff Hall, David Hogg, and Owen Johnson. Process mining in oncology using the mimic-iii dataset. In Journal of Physics: Conference Series, volume 971, page 012008. IOP Publishing, 2018.
13. Fuhai Li, Hui Xin, Jidong Zhang, Mingqiang Fu, Jingmin Zhou, and Zhexun Lian. Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the mimic-iii database. BMJ open, 11(7):page 044779, 2021.
14. Caijie Liu, Shuying Wang, and Xiuzhen Wang. Effect of transthoracic echocardiography on short-term outcomes in patients with acute kidney injury in the intensive care unit: a retrospective cohort study based on the mimic-iii database. Annals of translational medicine, 10(15), 2022.
15. Dongliang Liu, Yiyang Tang, and Qian Zhang. Admission hyperglycemia predicts long-term mortality in critically ill patients with subarachnoid hemorrhage: a retrospective analysis of the mimic-iii database. Frontiers in Neurology, 12:678998, 2021.
16. Vincent Liu, Gabriel J Escobar, John D Greene, Jay Soule, Alan Whippy, Derek C Angus, and Theodore J Iwashyna. Hospital deaths in patients with sepsis from 2 independent cohorts. Jama, 312(1):90–92, 2014.
17. Guiqiang Miao, Zhaohui Li, Linjian Chen,Wenyong Li, Guobo Lan, Qiyuan Chen, Zhen Luo, Ruijia Liu, and Xiaodong Zhao. A novel nomogram for predicting morbidity risk in patients with secondary malignant neoplasm of bone and bone marrow: an analysis based on the large mimic-iii clinical database. International Journal of General Medicine, pages 3255–3264, 2022.
18. Edward Moseley, Leo Anthony Celi, JoyWu, and Franck Dernoncourt. Phenotype annotations for patient notes in the mimic-iii database. PhysioNet, 2020.
19. Matthieu Scherpf, Felix Gr¨aser, Hagen Malberg, and Sebastian Zaunseder. Predicting sepsis with a recurrent neural network using the mimic iii database. Computers in biology and medicine, 113:103395, 2019.
20. Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, CraigMCoopersmith, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). Jama, 315(8):801–810, 2016.
21. Shan Yang, Lirui Cao, Yongfang Zhou, and Chenggong Hu. A retrospective cohort study: Predicting 90-day mortality for icu trauma patients with a machine learning algorithm using xgboost using mimic-iii database. Journal of Multidisciplinary Healthcare, pages 2625–2640, 2023.
22. Yiran Zhang, Qi Zheng, Xiaoyi Dai, Xingjie Xu, and Liang Ma. Overweight is associated with better one-year survival in elderly patients after cardiac surgery: a retrospective analysis of the mimic-iii database. Journal of Thoracic Disease, 13(2):562, 2021.